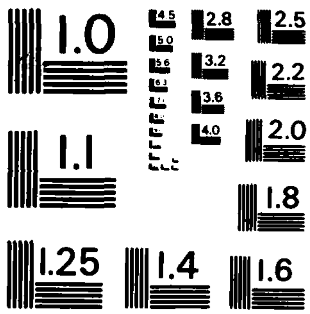
AD-A132 723 INFORMATIVE QUANTILE FUNCTIONS AND IDENTIFICATION OF PROBABILITY DISTRIBU..(U) TEXAS A AND M UNIV COLLEGE STATION DEPT OF STATISTICS E PARZEN AUG 83 TR-A-26 1/1

UNCLASSIFIED ARO-20140.5-MA DAAG29-83-K-0051 F/G 12/1 NL

END
DATE
FILMED
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3

ARO 20140.5-MA

(12)

TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS 77843-3143

STATISTICS
Phone 409 - 845-3141

AD-A132 723

INFORMATIVE QUANTILE FUNCTIONS AND IDENTIFICATION OF PROBABILITY DISTRIBUTION TYPES

by Emanuel Parzen

Department of Statistics

Texas A&M University

Technical Report No. A-26

August 1983

Texas A&M Research Foundation
Project No. 4858

"Functional Statistical Data Analysis and Modeling"

Sponsored by the Army Research Office

Contract DAAG29-83-K-0051

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER A-26	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Informative Quantile Functions and Identification of Probability Distribution Types		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s) DAAG29-83-K-0051
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Institute of Statistics College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE August 1983
		13. NUMBER OF PAGES 40
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Quantile data analysis, tail estimation, goodness of fit, exploratory data analysis.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number) A problem of great importance to statistical data analysts is quick identification of possible probability distributions for observed data, and classification of tail behavior of probability distributions. This paper discusses the informative quantile function $IQ(u) = \{Q(u) - Q(0.5)\} \div 2\{Q(0.75) - Q(0.25)\}$, and its use to identify probability models for observed data and its use to provide concepts of "representative distributions" which illustrate the different types of shapes and tail behavior that real distributions can have.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

INFORMATIVE QUANTILE FUNCTIONS AND
IDENTIFICATION OF PROBABILITY DISTRIBUTION TYPES

by Emanuel Parzen
Department of Statistics
Texas A&M University

Abstract

A problem of great importance to statistical data analysts is quick identification of possible probability distributions for observed data, and classification of tail behavior of probability distributions. This paper discusses the informative quantile function $IQ(u) = \{Q(u) - Q(0.5)\} \div 2\{Q(0.75) - Q(0.25)\}$, and its use to identify probability models for observed data and its use to provide concepts of "representative distributions" which illustrate the different types of shapes and tail behavior that real distributions can have. This paper also discusses estimators of tail exponents; they can be used to identify outlying data values, and more centrally to identify possible distributions to fit to data.

CONTENTS

0. Prologue: keys, two-keys, and statistical signals

This paper introduces the informative quantile function; its definition is probability based, its properties can be studied both mathematically and empirically, and it provides unified definitions and practical estimators of the tail types of probability distributions that can fit an observed batch of data. Illustrative tables of tail values of informative quantile functions of familiar distributions are given; they provide new types of keys (and two-keys) for exploratory data analysis of a (random) sample (of a random variable).

A key for exploratory data analysis is defined to be a method of data detection by which researchers can familiarize ourselves "with the data, get a rough idea of potential problems, and look for both obvious and subtle clues about the process that generated the data and the process that processed the data before we got to see it" [Welsch commenting on Parzen (1979)]. When a key is based on concepts of probability theory (and thus ultimately also provides methods of data inference and confirmatory data anlaysis), we call it a two-key.

Keys which are also two-keys provide statistical signals. One important role of numerical statistical signals is to be appended to statistical graphics to help guide the Viewer's attention to the graphical statistical signals (significant features of the graphs). In support of the proposition that the best keys are two-keys, we conclude with a statement by W. E. Deming entitled "Statistical Work and Computers." (We do not know where it was published, and believe it to have been written in the early 1970's).

> The feature that distinguishes the statistician from other professions is his use of the theory of probability. The statistician requires knowledge of statistical theory. To fulfill his duties in professional practice, he must distinguish between knowledge and wisdom. He is a scientist, but also an artist. He requires wisdom to make a good choice of problem and a choice of statistical procedure that will be valid and feasible under the circumstances.
>
> The computer can be the statistician's servant, though many people are content if it is the other way around. Many firms today have magnificent information systems, but too often these systems fail to present information as wisdom. The statistician, in his aim to find causes of variation in product (synonymous with poor quality and high costs), may use data from an information system, but he adapts the system to calculate statistical signals. It is more important to have a system to improve performance than to have a system that merely tells us where we are now. The statistician transforms information into a living force for the advancement of knowledge and for improvement of quality and output, industrial and agricultural.

1. Quantile and sample quantile functions

Various aspects of the probability distribution of a random variable X are described by its:

distribution function	$F(x) = \Pr[X \le x]$,	$-\infty < x < \infty$;
probability density	$f(x) = F'(x)$,	$-\infty < x < \infty$;
quantile function	$Q(u) = F^{-1}(u)$,	$0 \le u \le 1$;
quantile density function	$q(u) = Q'(u)$,	$0 \le u \le 1$;
density-quantile function	$fQ(u) = fF^{-1}(u)) = \{q(u)\}^{-1}$,	$0 \le u \le 1$;
score function	$J(u) = -(fQ)'(u)$,	$0 \le u \le 1$.

Let $X_1, X_2, \ldots, X_n$ be a data set. The keys we propose, to gain insight into the processes generating the data, become two-keys when we assume that the data batch is a random sample of a random variable X. The sample distribution function $\tilde{F}(x)$ and sample quantile function $\tilde{Q}(u)$ are defined in terms of the order statistics $X_{1n} \le X_{2n} \le \ldots \le X_{nn}$ of the sample:

$$\tilde{F}(x) = \frac{j}{n} \quad , X_{jn} \le x < X_{(j+1)n} \qquad ;$$

$$\tilde{Q}(u) = X_{jn}, \quad \frac{j-1}{n} < u \le \frac{j}{n} \qquad .$$

<u>In practice we prefer to use a sample quantile function</u> $\tilde{Q}(u)$ <u>which is piecewise linear between the values</u>

$$\tilde{Q}\left(\frac{j}{n+1}\right) = X_{jn} \quad , \; j=1,\ldots,n.$$

For graphical data analysis, we transform $\tilde{Q}(u)$ to a normalized version $I\tilde{Q}(u)$, called the sample informative quantile function. The value of $I\tilde{Q}(u)$, as u tends to 0 and 1, provide diagnostic measures of the <u>type</u> of probability distribution. An important classification of "type" is in terms of tail exponents.

2. Tail Exponents Classification of Probability Laws

From extreme value theory, statisticians have long realized that it is useful to classify distributions according to their tail behavior (behavior of $F(x)$ as x tends to $\pm\infty$). It is usual to distinguish three main types of distributions, called (1) limited, (2) exponential, and (3) algebraic. This classification can also be expressed in terms of the density quantile function $fQ(u)$; we call the types short, medium, and long tail.

A reasonable assumption about the distributions that occur in practice is that their density-quantile functions are regularly varying in the sense that there exist tail exponents α_0 and α_1 such that, as $u \to 0$,

$$fQ(u) = u^{\alpha_0} L_0(u) \quad , \quad fQ(1-u) = u^{\alpha_1} L_1(u)$$

where $L_j(u)$ for $j=0,1$ is a slowly varying function.

A function $L(u)$, $0<u<1$ is usually defined to be slowly varying as $u \to 0$ if, for every y in $0<y\leq 1$, $L(yu)/L(u) \to 1$ or $\log L(yu) - \log L(u) \to 0$. For estimation of tail exponents we will require further that, as $u \to 0$,

$$\int_0^1 \{\log L(yu) - \log L(u)\}\, dy \to 0$$

which we call integrally slowly varying. An example of a slowly varying function is $L(u) = \{\log u^{-1}\}^{\beta}$; this is proved in section 9.

<u>Classification of tail behavior of probability laws</u>

A probability law has a left tail type and a r'ght tail type depending on the value of α_0 and α_1. If α is the tail exponent, we define:

$\alpha < 0$	super short tail
$0 \le \alpha < 1$	short tail
$\alpha = 1$	medium tail
$\alpha > 1$	long tail

Medium tailed distributions are further classified by the value of $J^* = \lim J(u)$:

$\alpha = 1$,	$J^* = 0$	medium long tail
$\alpha = 1$,	$0 < J^* < \infty$	medium-medium tail
$\alpha = 1$,	$J^* = \infty$	medium-short tail

One immediate insight into the meaning of tail behavior is provided by the hazard function

$$h(x) = f(x) \div \{1-F(x)\}$$

with hazard quantile function $hQ(u) = fQ(u) \div 1-u$. The convergence behavior of $h(x)$ as $x\to\infty$ is the same as that of $hQ(u)$ as $u\to 1$. From the definitions one sees that $h^* = \lim_{x\to\infty} h(x)$ satisfies

$h^* = \infty$	(increasing hazard rate)	Short or medium-short tail
$0<h^*<\infty$	(constant hazard rate)	Medium-medium tail
$h^* = 0$	(decreasing hazard rate)	Long or medium-long tail

3. Unitized and Informative Quantile Functions

If one can define "universal" location and scale parameters, denoted μ_1 and σ_1 respectively, then one can define a normalization of the quantile function which depends only on its shape (and is independent of location and scale) by

$$Q_1(u) = \frac{Q(u) - \mu_1}{\sigma_1} \quad .$$

We propose

$$\mu_1 = Q(0.5), \qquad \sigma_1 = Q'(0.5) = q(0.5) \quad .$$

We call $Q_1(u)$ the unitized quantile function.

One can distinguish three kinds of estimators of parameters [such as μ_1 and σ_1]: fully non-parametric [denoted $\tilde{\mu}_1$ and $\tilde{\sigma}_1$], fully parametric [denoted $\hat{\mu}_1$ and $\hat{\sigma}_1$], and functional [estimators $\breve{\mu}_1$ and $\breve{\sigma}_1$ which are the parameters of smoothed quantile functions $\breve{Q}(u)$ obtained by smoothing the raw or fully non-parametric estimator $\tilde{Q}(u)$]. The shape of $Q(u)$ must be inferred before one can efficiently estimate μ and σ using fully parametric (or robust parametric) estimators.

A fully non-parametric estimator of $Q(0.5)$ is $\tilde{Q}(0.5)$. A fully non-parametric estimator of $q(0.5)$ is more difficult to define. We therefore consider quick and dirty approximators of $q(0.5)$ of the form

$$\sigma_p = \frac{Q(0.5 + p) - Q(0.5 - p)}{2p}$$

where $0 \leq p \leq 0.5$. We usually take $p = 0.25$; then we approximate $q(0.5)$ by

$$\sigma_{0.25} = 2\{Q(0.75) - Q(0.25)\} \quad .$$

We call

$$IQ(u) = \frac{Q(u) - Q(0.5)}{2\{Q(0.75) - Q(0.25)\}}$$

the <u>informative</u> <u>quantile</u> <u>function</u>.

We compute $IQ(u)$, but graphically we plot the <u>truncated</u> <u>informative</u> <u>quantile</u> function

$$\begin{aligned} TIQ(u) &= -1 \text{ if } IQ(u) < -1, \\ &= 1 \text{ if } IQ(u) > 1, \\ &= IQ(u) \text{ if } |IQ(u)| \leq 1. \end{aligned}$$

In addition to the plot of $TIQ(u)$, we report the values of $IQ(u)$ at u=0.01, 0.05, 0.10, 0.25, 0.75, 0.90, 0.95, 0.99. Truncating the values of $IQ(u)$ in our plot enables us to see the "middle" of the distribution. The ends (tails) of the distributions are described numerically by the extreme values of $IQ(u)$.

For convenience in seeing at a glance in a plot of IQ(u) its behavior, especially as u tends to 0 and 1, we plot on the same graph the IQ(u) of a uniform distribution (it is a straight line with values -0.5 and 0.5 at u = 0 and 1 respectively).

Example: Super Short Distributions. An imporant example of a super-short distribution ($\alpha<0$) is $X = -\cos \pi U$ where U is uniform [0,1]. Since $-\cos \pi u$ is an increasing function of u, the quantile function of X is $Q(u) = -\cos \pi u$, with quantile density and density-quantile

$$q(u) = \frac{\sin \pi u}{\pi} \quad , \qquad fQ(u) = \frac{\pi}{\sin \pi u} \quad .$$

As $u \to 0$, $fQ(u) \sim u^{-1}$ so $\alpha_0 = -1$. The distribution is symmetric, in the sense that $q(1-u) = q(u)$; therefore $\alpha_1 = -1$. The interquartile range $IQR = \sqrt{2}$; the informative quantile function is $IQ(u) = (-.35) \cos \pi u$. Therefore $IQ(0) = -.35$, $IQ(1) = .35$. These values are taken as typical values of super-short distributions.

4. Examples of theoretical informative quantile functions

A normal distribution is defined in terms of the standard normal density $\phi(x)$ and distribution $\Phi(x)$,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2} x^2 \ , \quad \Phi(x) = \int_{-\infty}^{\infty} \phi(y)\,dy;$$

a distribution $F(x)$ is called normal when it can be represented

$$F(x) = (\frac{x-\mu}{\sigma}) \ , \quad f(x) = \frac{1}{\sigma} (\frac{X-\mu}{\sigma})$$

with quantile function

$$Q(u) = \mu + \sigma \, \Phi^{-1}(u).$$

The parameters μ_1 and σ_1 are related to μ and σ by $\mu_1 = \mu$ and $\sigma_1 = \sigma\sqrt{2\pi}$. The unitized normal density (for which $\sigma_1 = 1$) has density

$$f_1(x) = \sqrt{2\pi} \ \ \phi(x \sqrt{2\pi}) = e^{-\pi x^2}$$

which is Stigler's proposal for a standardized normal density [Stigler (1982)].

An exponential distribution has density

$$f(x) = \frac{1}{\sigma} f_o(\frac{x}{\sigma}) \ , \quad f_o(x) = e^{-x} \ , \quad x \geq 0$$

and quantile function

$$Q(u) = \log (1-u)^{-1} \quad .$$

Although its mean equals σ, we regard σ as a scale parameter rather than a location parameter. The parameters μ_1, σ_1, and $\sigma_{0.25}$ satisfy

$$\mu_1 = \sigma \log 2 = (.69)\ \sigma; \quad \sigma_1 = 2\sigma \quad ; \quad \sigma_{0.25} = 2.2\sigma \quad .$$

The unitized and informative exponential quantile functions are

$$Q_1(u) = -0.5 \log 2(1-u)$$

$$IQ(u) = -0.45 \log 2(1-u) \quad .$$

The possible shapes of informative quantile functions are best described by plots of the Weibull distribution with parameter β, which has standard quantile function

$$Q(u) = \{\log (1-u)^{-1}\}^{\beta} \quad .$$

Graphs of the information quantile functions of the Weibull distribution for β = .1 (.1) 2.0 are given in the appendix.

5. Outlying data value interpretation of $\tilde{IQ}(u)$

The sample informative quantile function is defined by

$$\tilde{IQ}(u) = \{\tilde{Q}(u) - \tilde{Q}(0.5)\} \div 2\ \tilde{IQR}$$

where $\tilde{IQR}$ is the sample interquartile range: $\tilde{IQR} = \tilde{Q}(0.75) - \tilde{Q}(0.25)$. The truncated sample informative quantile function $\tilde{TIQ}(u)$ is defined to be $\tilde{IQ}(u)$ truncated at ± 1.

Hoaglin, Mosteller, and Tukey (1983, p. 39) introduce techniques for identifying outlying (or outside) data values as those lying outside the interval

$$(\tilde{Q}(0.25) - (1.5)\ IQR,\ \ \tilde{Q}(0.75) + (1.5)\ IQR) \quad .$$

We regard as outlying data values those lying outside the interval

$$(\tilde{Q}(0.5) - 2\tilde{IQR},\quad \tilde{Q}(0.5) + 2\ \tilde{IQR}) \quad .$$

Outlying data values appear on the plot of $\tilde{TIQ}(u)$ as values truncated to ± 1. The actual values of outlying data values are represented by the values of $\tilde{IQ}(u)$ for u=0.01, 0.05, 0.10, 0.90, 0.95, 0.99. The next section discusses how these quantities provide quick and dirty estimators of the tail type of the distributions that can fit the sample.

Other useful numerical diagnostics are estimators of the IQ-mean μIQ and IQ-standard-deviation σIQ, defined by

$$\mu IQ = \frac{\mu - \mu_1}{\sigma_{.25}} \quad , \quad \sigma IQ = \frac{\sigma}{\sigma_{.25}}$$

where μ and σ^2 are the mean and variance of Q(u). The logarithm (to the base e) of σID is denoted log SDIQ. For a normal distribution $\sigma ID = 1/27$ and log SDIQ = -1 approximately. A test that the sample has a Gaussian distribution can be based on testing if the sample estimator of log SDIQ is significantly different from -1.

6. Tables of tail values of informative quantile functions

One use of the informative quantile function $\tilde{IQ}(u)$ of a sample is to determine quickly probability distribution that might fit the sample. One can readily distinguish whether the data could be fit by a normal distribution or an exponential distribution [and thus determine the "probability of success" if one were to apply a more formal goodness of fit test]. However no standard parametric model may fit the data, and statistical data analysis must identify significant features of the data "non-parametrically".

Statistical scientists are seeking to define concepts which illustrate the different types of shapes and tail behavior that real distributions can have. Hoaglin, Mosteller, and Tukey (1983, p. 316) use language such as "neutral tailed (Gaussian)" and stretch-tailed (Cauchy)". To describe the notion of tail weight, they write that it "expresses how the extreme portion of the distribution spreads out relative to the width of the center." As an index of tail behavior, they introduce (p. 323)

$$\{\tilde{Q}(0.9) - \tilde{Q}(0.1)\} \div \{\tilde{Q}(0.75) - \tilde{Q}(0.25)\} = 2\{\tilde{IQ}(0.9) - \tilde{IQ}(0.1)\} .$$

As indices of tail behavior, this paper proposes $\tilde{IQ}(u)$ at $u = 0.01, 0.05, 0.1, 0.9, 0.95, 0.99$. The true values of these indices for various familiar distributions are given in the tables. These indices are keys (useful for exploratory data analysis of what's unusual or extraordinary about a data set) and two-keys (provide estimates of the tail exponents and tail types of distributions that might have generated the data).

Table 6A

Tail Values of Informative Quantile Function IQ(u)

Standard Distributions

* = Approximate value of u at which IQ(u) = 1.

Distribution	*	u .01	.05	.10	.90	.95	.99
Normal	--	-.862	-.610	-.475	.475	.610	.862
Exponential	.95	-.311	-.292	-.268	.732	1.048	1.780
Logistic	.99	-1.046	-.670	-.500	.500	.670	1.046
Double Exp	.97	-1.411	-.830	-.568	.580	.830	1.411
Cauchy	.92	-7.955	-1.578	-.769	.769	1.578	7.954
Extreme Value	--	-1.346	-.828	-.599	.382	.465	0.602
Log Normal	.91	-.310	-.278	-.278	.895	1.438	3.178
Super Short	--	-.353	-.349	-.336	.336	.349	0.353

Table 6B

Tail Values of Informative Quantile Function IQ(u)

Weibull $Q(u) = \{\log (1-u)^{-1}\}^{\beta}$

* = Approximate value of u at which IQ(u) = 1.

β	*	u= .01	.05	.10	.90	.95	.99
.1	--	-1.107	-.735	-.550	.409	.505	.668
.2	--	-.921	-.655	-.506	.438	.549	.743
.3	--	-.777	-.585	-.466	.468	.595	.826
.4	--	-.662	-.525	-.430	.500	.646	.919
.5	1.0	-.571	-.473	-.396	.534	.701	1.024
.6	.98	-.498	-.427	-.366	.570	.760	1.142
.7	.97	-.437	-.387	-.338	.607	.824	1.275
.8	.96	-.388	-.351	-.312	.647	.893	1.424
.9	.95	-.346	-.320	-.295	.689	.967	1.592
1.0	.94	-.311	-.292	-.273	.732	1.048	1.780
1.1	.93	-.281	-.267	-.252	.778	1.135	1.993
1.2	.93	-.255	-.245	-.233	.827	1.229	2.232
1.3	.92	-.232	-.225	-.216	.878	1.331	2.502
1.4	.91	-.212	-.207	-.200	.931	1.440	2.806
1.5	.90	-.195	-.191	-.185	.987	1.559	3.148
1.6	.89	-.179	-.177	-.172	1.046	1.687	3.54
1.7	.89	-.165	-.163	-.159	1.107	1.825	3.969
1.8	.88	-.153	-.151	-.147	1.172	1.974	4.459
1.9	.88	-.141	-.140	-.137	1.240	2.135	5.012
2.0	.87	-.131	-.130	-.128	1.311	2.309	5.635
2.1	.87	-.121	-.121	-.119	1.386	2.497	6.338
2.2	.86	-.112	-.112	-.111	1.464	2.700	7.130
2.3	.86	-.104	-.104	-.103	1.546	2.919	8.023
2.4	.85	-.097	-.097	-.096	1.633	3.155	9.031

Table 6C

Tail Values of Informative Quantile Function IQ(u)

Lognormal $Q(u) = \exp \lambda\Phi^{-1}(u)$

* = Approximate value of u at which IQ(u) = 1.

λ	*	u= .01	.05	.10	.90	.95	.99
.5	.96	-.500	-.408	-.344	.653	.928	1.600
1	.92	-.310	-.278	-.246	.895	1.438	3.178
1.5	.88	-.203	-.192	-.179	1.223	2.260	6.655
2	.86	-.138	-.134	-.128	1.666	3.594	14.449
2.5	.84	-.096	-.094	-.092	2.266	5.761	32.083
3	.82	-.067	-.067	-.066	3.077	9.284	72.169
3.5	.81	-.048	-.047	-.047	4.175	15.012	163.511
4	.80	-.034	-.034	-.034	5.661	24.322	371.888
4.5	.80	-.024	-.024	-.024	7.673	39.454	847.538
5	.79	-.017	-.017	-.017	10.398	64.041	--
5.5	.79	-.012	-.012	-.012	14.089	103.988	--
6	.79	-.009	-.009	-.009	19.087	168.886	--
6.5	.78	-.006	-.006	-.006	25.858	274.315	--
7	.78	-.004	-.004	-.004	35.029	445.586	--
7.5	.78	-.003	-.003	-.003	47.452	723.814	--
8	.78	-.002	-.002	-.002	64.280	--	--

7. Example of sample informative quantile analysis

A data set extensively analyzed at Bell Telephone Laboratories (and discussed in a recent book on graphical methods of data analysis by Chambers, Cleveland, Kleiner, and Tukey, (1983)) consists of Stamford Conn. Monthly Maximum Ozone levels. Sample size n=136, sample median $\tilde{\mu}_1 = 80$, sample mean $\tilde{\mu} = 89.7$, twice interquartile range $\tilde{\sigma}_1 = 147.5$, and standard deviation $\tilde{\sigma} = 52.1$. Rather than reporting the original data $X_1,\ldots,X_n$ we report (table 7A) the normalized values $(X_j-\tilde{\mu}_1) \div \tilde{\sigma}_1$ which are used to plot $\tilde{IQ}(u)$; a plot of $\tilde{Q}(u)$ is given on p. 15 of Chambers et al. Numerical statistical signals are provided by the tail values:

u	0.05	.1	.90	.95
$\tilde{IQ}(u)$	-.38	-.33	.61	.83

By consulting the table of Weibull informative quantile values, as a first guess of a distribution to fit this data one takes Weibull with parameter $\beta = 0.8$. The graph of $\tilde{IQ}(u)$ in Figure 7A also suggests to us that a Weibull distribution provides a good first approximation. How to refine this approximation is a problem treated by our ONESAM data analysis program.

An alternate approach to modeling this data is to find a transformation to normality; one would then report as one's conclusion that cube root of Stamford Ozone data is normally distributed. We believe that this conclusion must be considered curve fitting, while a conclusion that the data is fit by a

Weibull distribution with β in a specified range represents a curve fit with scientific insight (which may help to explain the physical mechanisms generating the data).

FIGURE 7A

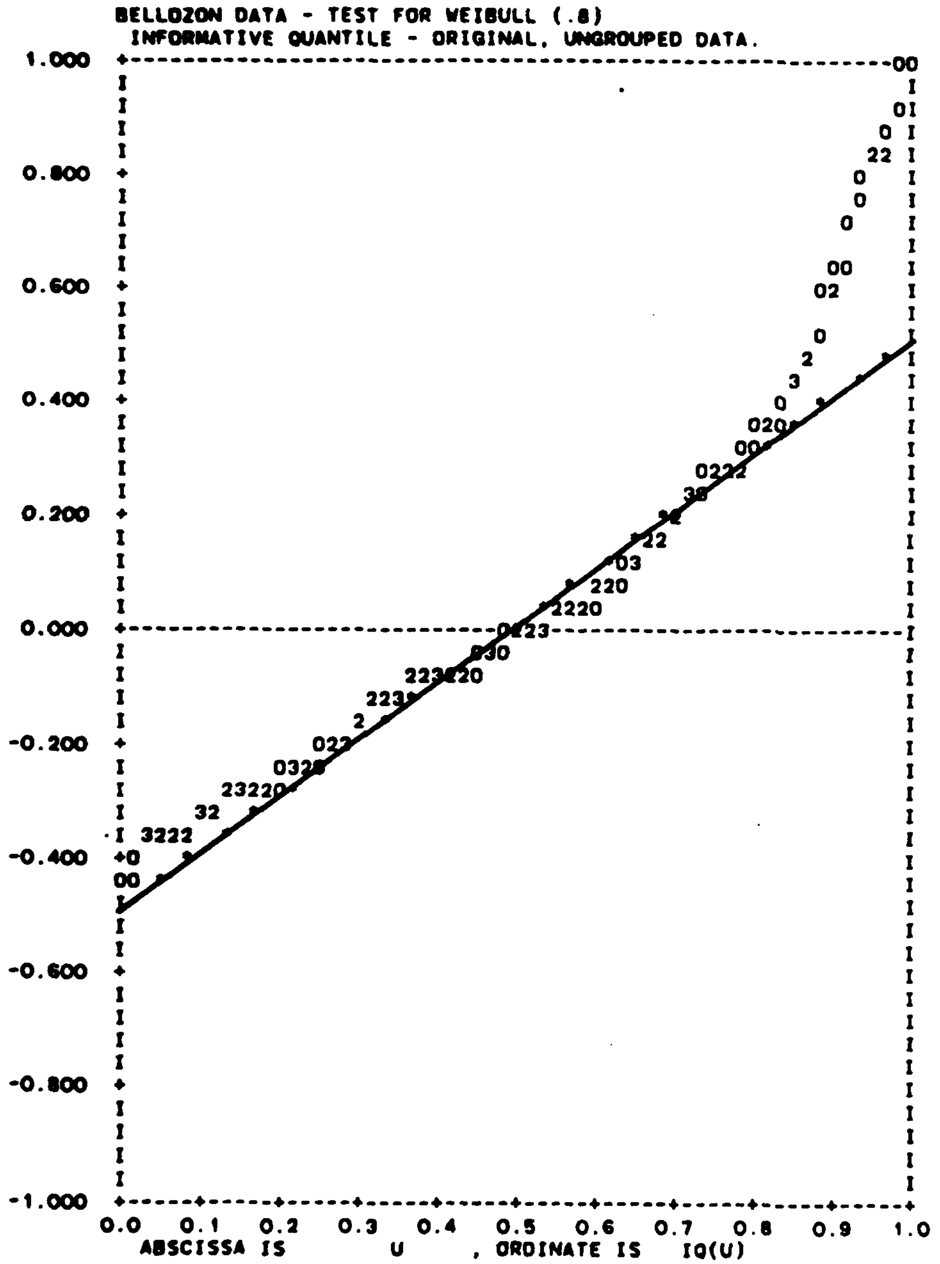

BELLOZON DATA - TEST FOR WEIBULL (.8)
INFORMATIVE QUANTILE - ORIGINAL, UNGROUPED DATA.
1.000
0.800
0.600
0.400
0.200
0.000
-0.200
-0.400
-0.600
-0.800
-1.000
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
ABSCISSA IS U , ORDINATE IS IQ(U)

TABLE 7A

BELLOZON DATA - TEST FOR WEIBULL (.8)
INFORMATIVE QUANTILE - ORIGINAL, UNGROUPED DATA.

ORDER STATISTICS IN QUARTERS

SEQUENCE WITHIN QUARTILE	FIRST QUARTER	SECOND QUARTER	THIRD QUARTER	FOURTH QUARTER
1	-0.4475	-0.2102	0.0	0.2847
2	-0.4475	-0.1966	0.0	0.2847
3	-0.3864	-0.1898	0.0	0.2983
4	-0.3797	-0.1898	0.0	0.2983
5	-0.3797	-0.1898	0.0136	0.2983
6	-0.3797	-0.1898	0.0136	0.3051
7	-0.3797	-0.1695	0.0203	0.3051
8	-0.3661	-0.1424	0.0339	0.3458
9	-0.3593	-0.1356	0.0407	0.3593
10	-0.3525	-0.1288	0.0407	0.3661
11	-0.3525	-0.1288	0.0475	0.3797
12	-0.3525	-0.1085	0.0475	0.4136
13	-0.3322	-0.1085	0.0475	0.4203
14	-0.3322	-0.1085	0.0610	0.4271
15	-0.3254	-0.1085	0.0746	0.4475
16	-0.3254	-0.0949	0.0814	0.4746
17	-0.3186	-0.0949	0.0949	0.4881
18	-0.2915	-0.0814	0.0949	0.5085
19	-0.2847	-0.0814	0.1220	0.6034
20	-0.2847	-0.0814	0.1288	0.6034
21	-0.2847	-0.0746	0.1288	0.6102
22	-0.2847	-0.0610	0.1356	0.6305
23	-0.2847	-0.0610	0.1424	0.6373
24	-0.2847	-0.0610	0.1559	0.7322
25	-0.2847	-0.0610	0.1559	0.7593
26	-0.2847	-0.0610	0.1559	0.7864
27	-0.2712	-0.0610	0.1898	0.8203
28	-0.2576	-0.0542	0.2102	0.8271
29	-0.2508	-0.0542	0.2237	0.8271
30	-0.2305	-0.0475	0.2237	0.8542
31	-0.2237	-0.0339	0.2305	0.8949
32	-0.2237	-0.0339	0.2576	0.9153
33	-0.2237	0.0	0.2644	1.0169
34	-0.2237	0.0	0.2644	1.0847

8. Super-short distributions as harbingers of bimodality

When the sample informative quantile function indicates a "super short" distribution the true distribution may not be a super-short unimodal distribution, but a bimodal distribution.

The manner in which a super-short distribution may be indicative of bimodality is indicated by the two-sample problem. One has a sample of values from a distribution $F(x)$, and a sample of values from a distribution $G(x)$. When the samples are pooled, they are regarded as a sample from a distribution $H(x)$ which can be represented $H(x) = \lambda F(x) + (1-\lambda) G(x)$ where λ is the fraction of the pooled sample from $F(x)$. One often seeks to test the hypothesis H_o: $F(x) = G(x)$. The informative quantile plot of $H(x)$ is super-short when F and G have their modes far apart.

To illustrate the ideas, assume $F(x) = \Phi(x)$, $G(x) = \Phi(x-\delta)$, $H(x) = 0.5\{\Phi(x) + \Phi(x-\delta)\}$. A random sample from $H(x)$, of size 40 was simulated, for $\delta = 1, 2, 3, 4, 5, 6$. The observed values of $\tilde{IQ}(u)$ are given in the following table.

δ u	.05	.10	.25	.75	.90	.95
1	-.6566	-.6069	-.2110	.2890	.5005	.6570
2	-.4450	-.3553	-.2044	.2956	.5847	.7258
3	-.4077	-.2801	-.2034	.2966	.5012	.6108
4	-.4586	-.4260	-.2908	.2092	.3326	.4340
5	-.4350	-.3620	-.2649	.2351	.4079	.4191
6	-.3228	-.2915	-.1841	.3159	.3795	.4179

Other summary statistics of the samples were

δ	Median	Interquartile Range	Mean IQ	St. Dev. IQ	Log SDIQ
1	.62	1.46	.01	.3689	-.997
2	1.10	2.07	.05	.3347	-1.095
3	.97	2.85	.05	.3024	-1.196
4	2.23	3.96	-.03	.2846	-1.257
5	2.36	4.00	.01	.2900	-1.238
6	2.39	5.28	.05	.2669	-1.321

The values of $\tilde{IQ}(0.05)$, $\tilde{IQ}(0.95)$ and log SDIQ in the case $\delta = 1$ indicate a Gaussian distribution. The values of $\tilde{IQ}(0.05)$ and $\tilde{IQ}(0.95)$ in the cases $\delta = 4, 5, 6$ indicate a super-short distribution which leads us to check the quantile functions of the pooled sample for the possiblity of bimodality which often indicates that the two samples do not have the same distributions.

9. Theoretical and empirical formulas for computing tail exponents

The properties of slowly varying functions are best understood by considering an example.

Lemma $L(u) = \{\log u^{-1}\}^{\beta}$ is (integrally) slowly varying as $u \to 0$.

Proof: $\log L(yu) = \beta \log \log (yu)^{-1} = \beta \log \{\log y^{-1} + \log u^{-1}\}$.

$$\log L(yu) - \log L(u) = \beta \log \{1 + (\log y^{-1}/\log u^{-1})\}$$

$$|\log L(yu) - \log L(u)| \leq \beta \, |(\log y^{-1}/\log u^{-1}| \quad .$$

Verify that $\int_0^1 |\log y| \, dy < \infty$, and $1/\log u^{-1} \to 0$ as $u \to 0$. One can conclude that $L(u)$ is slowly varying and also integrally slowly varying.

The representation of $fQ(u)$ suggests a formula for computation of tail exponents α_0 and α_1 (which may be adapted to provide estimators from data).

Theorem: Computation of tail exponents

$$-\alpha_0 = \lim_{u \to 0} \int_0^1 \{\log fQ(yu) - \log fQ(u)\} \, dy \quad .$$

Equivalently

$$-\alpha_0 = \lim_{p \to 0} \frac{1}{p} \int_0^p \log fQ(t) \, dt - \log fQ(p) \quad .$$

Similarly

$$\alpha_1 = \lim_{u\to 0} \int_o^1 \{\log fQ(1-yu) - \log fQ(1-u)\}\, dy$$

$$= \lim_{p\to 1} \frac{1}{1-p} \int_p^1 \log fQ(t)\, dt - \log fQ(1-p) \quad .$$

Proof: $\log fQ(u) = \alpha_0 \log u + \log L_0(u)$,

$\log fQ(yu) - \log fQ(u) = \alpha_0 \log y + \log L_0(yu) - \log L_0(u$

Since $\int_o^1 \log y\, dy = -1$, we conclude that

$$\int_o^1 \{\log fQ(yu) - \log fQ(u)\}\, dy = -\alpha_0 + o(u) \quad .$$

Similarly one derives formula for α_1.

Because the density-quantile and quantile-density functions are reciprocals, we obtain similar formulas for $q(u)$ which may be easier to implement in practice:

$$q(u) = u^{-\alpha_0} L_0(u) \quad , \quad \text{as } u\to 0 \quad ,$$

$$q(u) = (1-u)^{-\alpha_1} L_1(1-u), \quad \text{as } u\to 1 \quad ;$$

$$\alpha_0 = \lim_{u\to 0} \int_o^1 \{\log q(yu) - \log q(u)\}\, dy \quad ;$$

$$\alpha_1 = \lim_{u\to 0} \int_o^1 \{\log q(1-yu) - \log q(1-u)\}\, dy.$$

For theoretical purposes it is often convenient to compute tail exponents using formulas such as

$$\alpha_0 = \lim_{u\to 0} u \frac{d}{du} \log fQ(u)$$

$$= \lim_{u\to 0} \frac{-u\, J(u)}{fQ(u)} \quad ;$$

$$\alpha_1 = \lim_{u\to 1} - (1-u) \frac{d}{du} \log fQ(u)$$

$$= \lim_{u\to 1} \frac{(1-u)\, J(u)}{fQ(u)} \quad .$$

In practice, we would estimate tail exponents from the values of $fQ(t)$ at an equispaced grid of points $t=j/n$, $j=1,2,\ldots,n-1$. Let k and n tend to ∞ in such a way that k/n tends to 0; define

$$-\alpha_{0,k} = \frac{1}{k} \sum_{j=1}^{k} \log fQ(\tfrac{j}{n}) - \log fQ(\tfrac{k+1}{n}) \quad ,$$

$$\alpha_{1,k} = \frac{1}{k} \sum_{j=n-k}^{n-1} \log fQ(\tfrac{j}{n}) - \log fQ(1-\tfrac{k+1}{n}) \quad .$$

Conjectures to be proved are that

$$\alpha_0 = \lim_{\substack{k\to\infty \\ k/n\to 0}} \alpha_{0,k}$$

$$\alpha_1 = \lim_{\substack{k\to\infty \\ k/n\to 0}} \alpha_{1,k} \quad .$$

The rate of convergence can be very slow. If $L(u) = \{\log u^{-1}\}^{\beta}$, then

$$\alpha_0 = \alpha_{0,k} + c\left|\log \frac{n}{k}\right|^{-1} \quad .$$

The theoretical properties and practical implementation of the foregoing estimators remains to be investigated. Related estimators are given in Mason (1982) and the papers referenced there.

References

Chambers, J. M., Cleveland, W. S., Kleiner, B., Tukey, P. A. (1983) Graphical Methods for Data Analysis, Duxbury: Boston.

Hoaglin, C., Mosteller, F. and Tukey, J. W. (1983) Understanding Robust and Exploratory Data Analysis, Wiley: New York.

Mason, D. M. (1982) Laws of large numbers for extreme values. Annals of Probability, 10, 754-764.

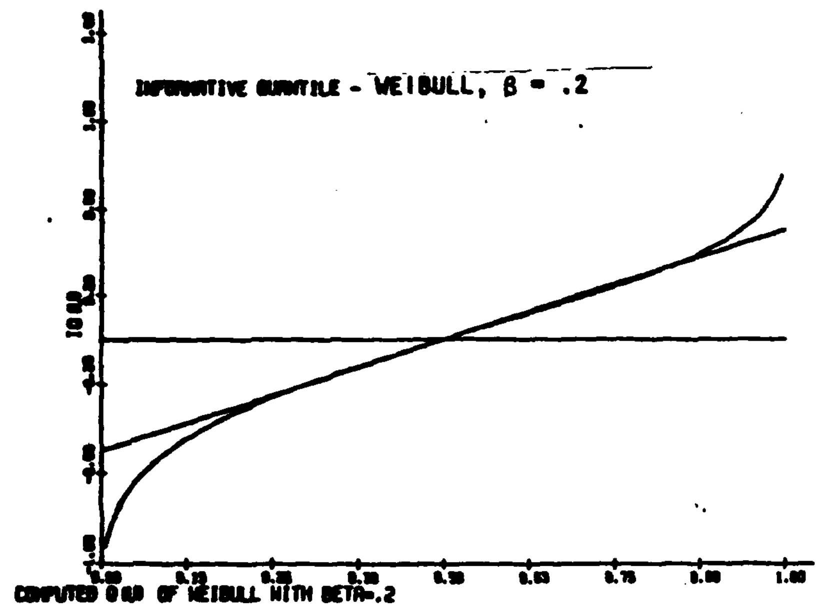
Parzen, E. (1979) Nonparametric Statistical Data Modeling. Journal of the American Statistical Association, 74, 105-131.

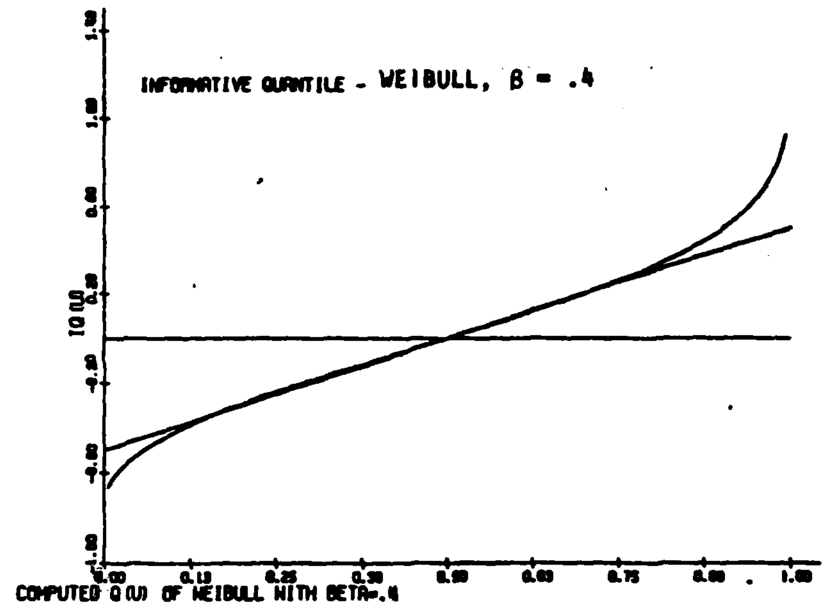
Stigler, S. M. (1982) A Modest Proposal: A New Standard for the Normal. The American Statistician, 36, 137-138.

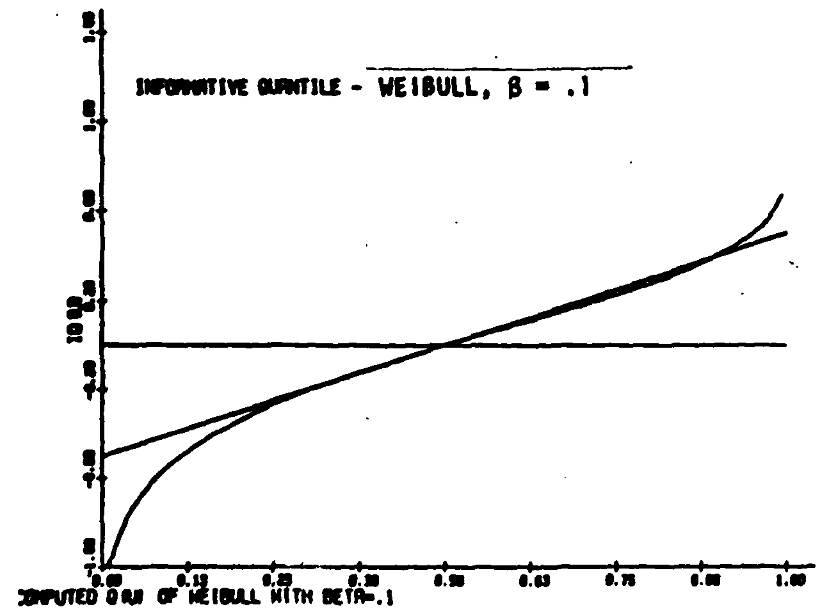
APPENDIX

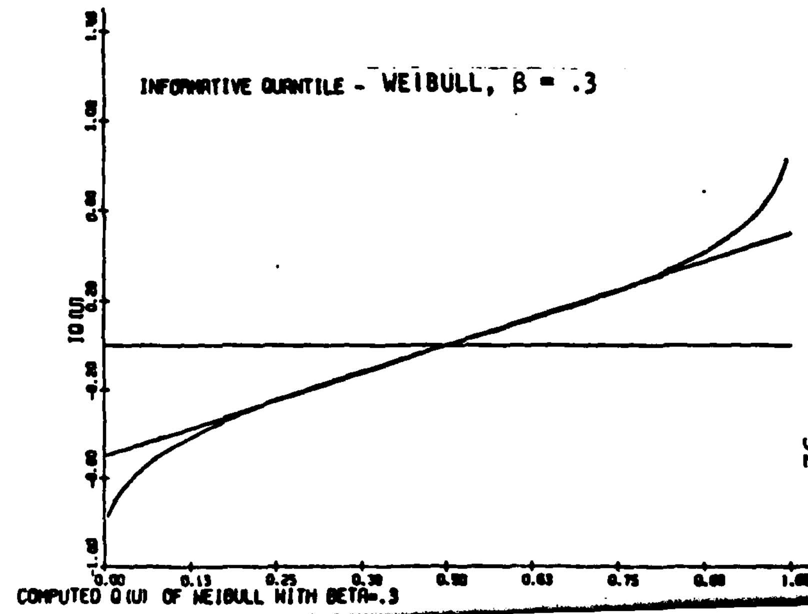
Informative Quantile Functions of Weibull Distributions with Parameter β:

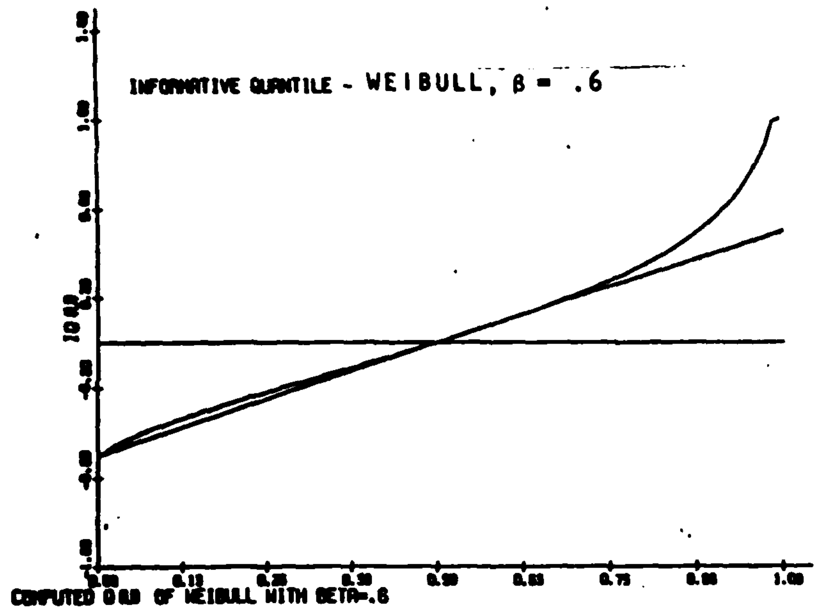
$$Q(u) = \{\log(1-u)^{-1}\}^{\beta}$$

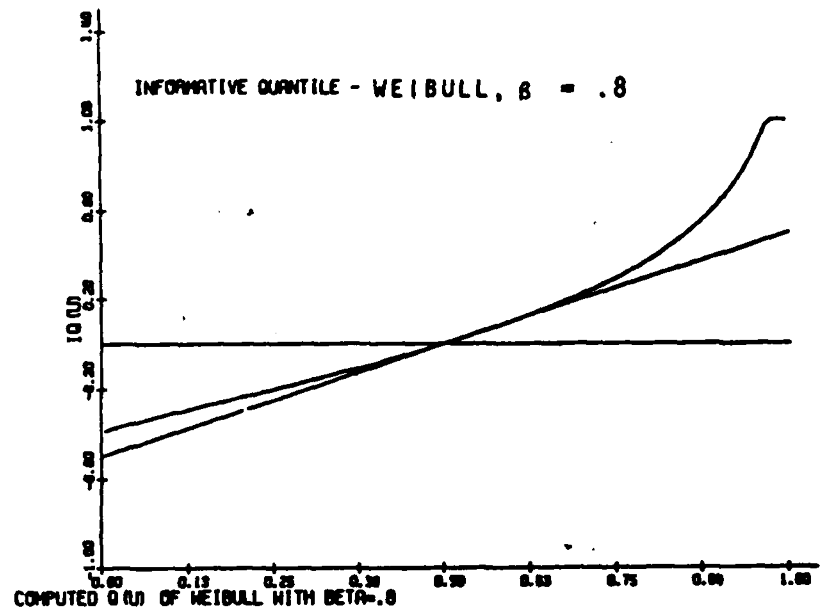

INFORMATIVE QUANTILE - WEIBULL, β = .2
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.2

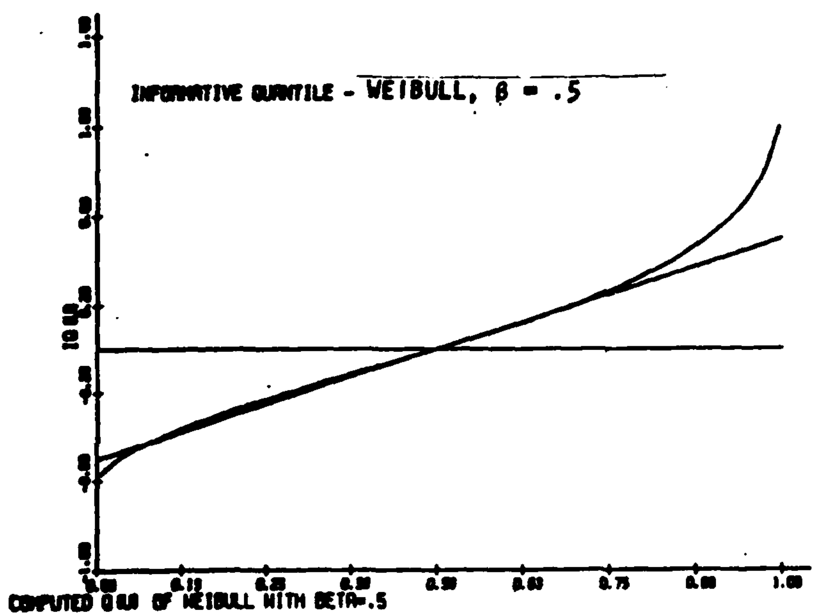

INFORMATIVE QUANTILE - WEIBULL, β = .4
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.4

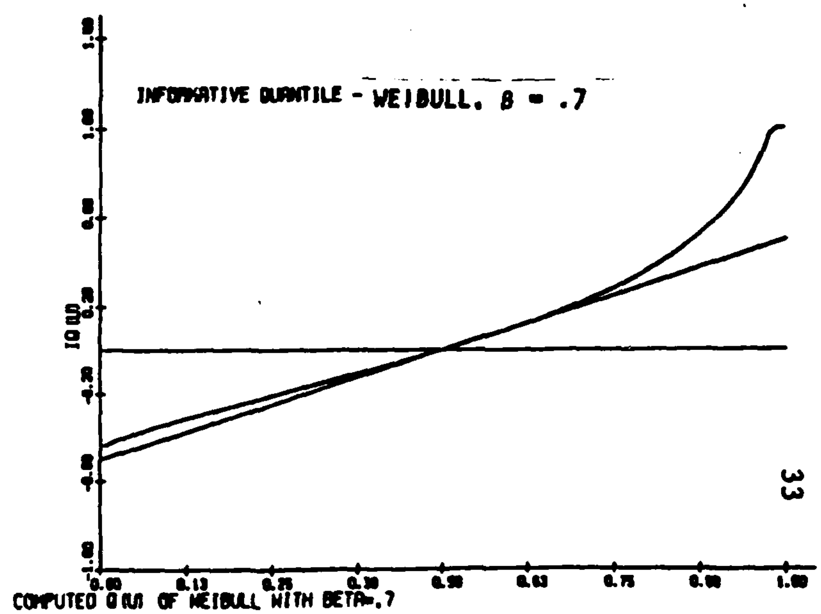

INFORMATIVE QUANTILE - WEIBULL, β = .1
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.1

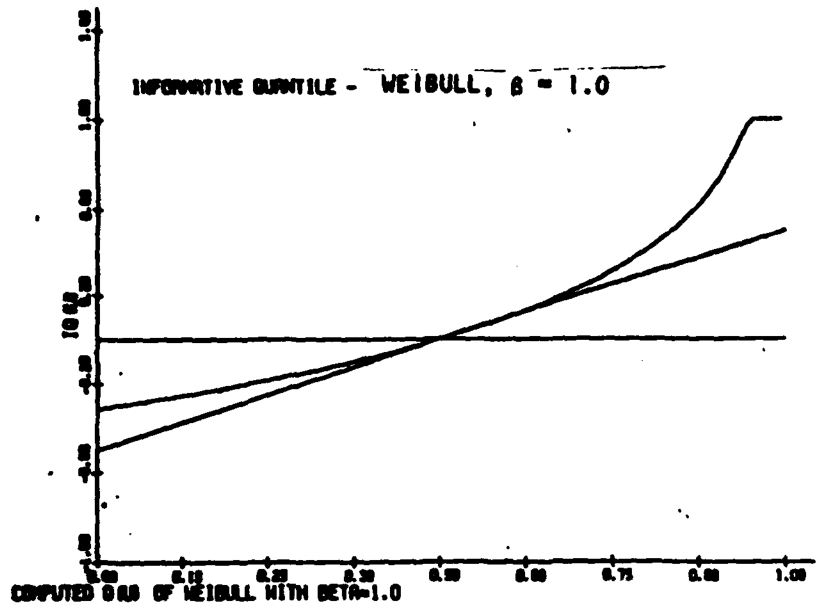

INFORMATIVE QUANTILE - WEIBULL, β = .3
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.3

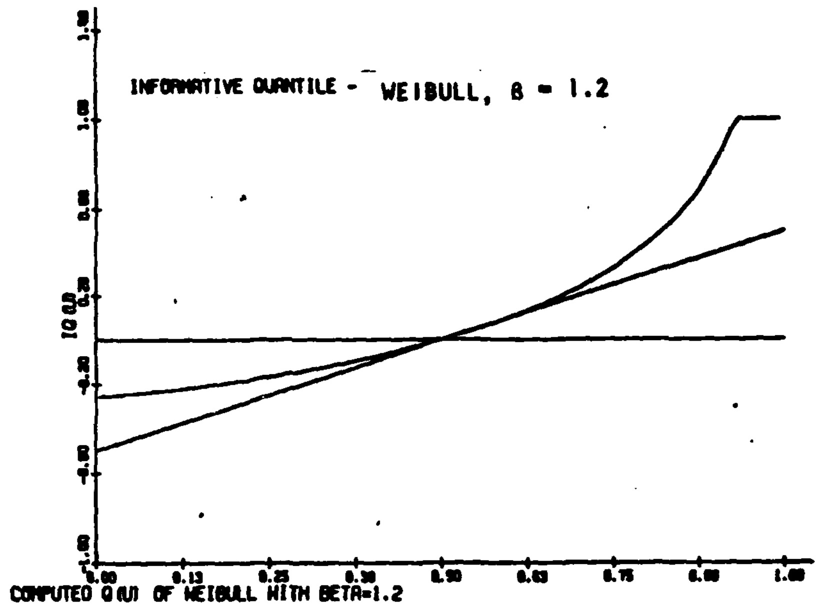

INFORMATIVE QUANTILE - WEIBULL, β = .6
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.6

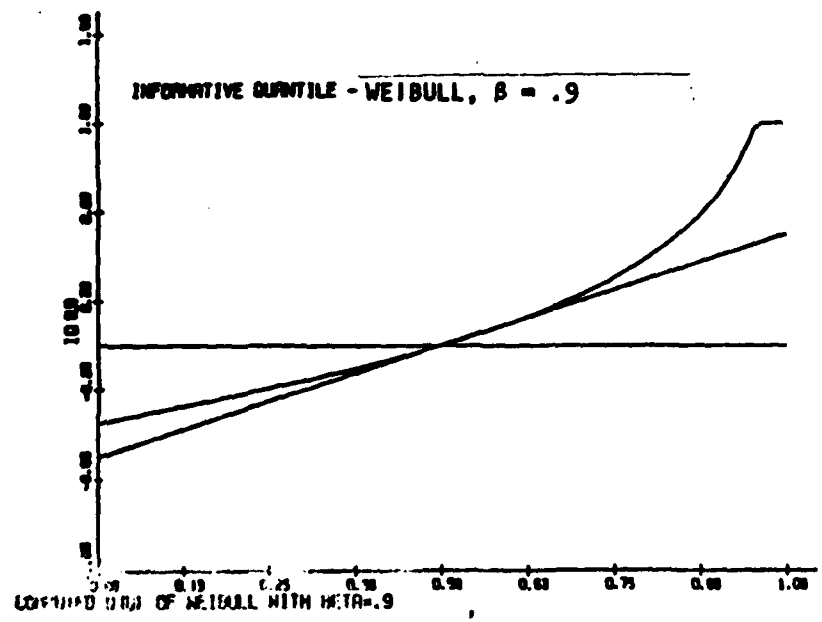

INFORMATIVE QUANTILE - WEIBULL, β = .8
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.8

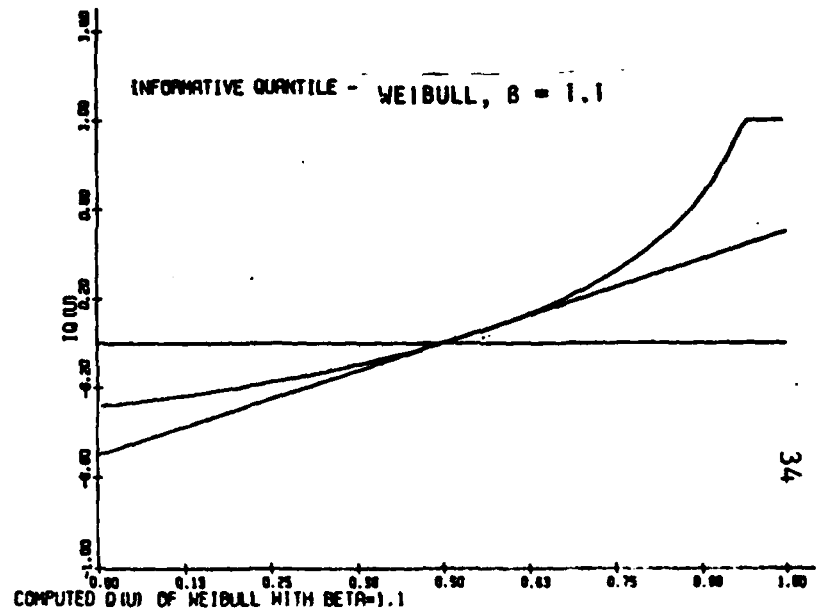

INFORMATIVE QUANTILE - WEIBULL, β = .5
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.5

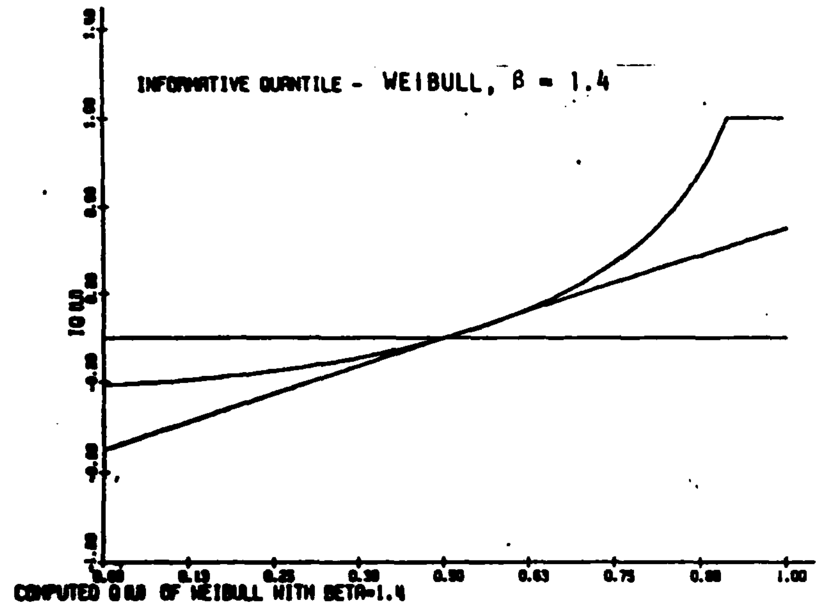

INFORMATIVE QUANTILE - WEIBULL, β = .7
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.7

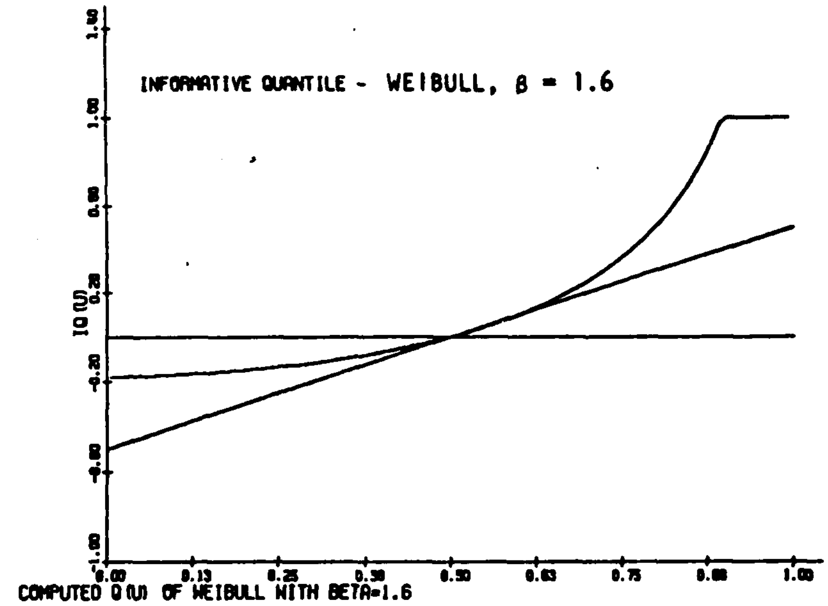

INFORMATIVE QUANTILE - WEIBULL, β = 1.0
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=1.0

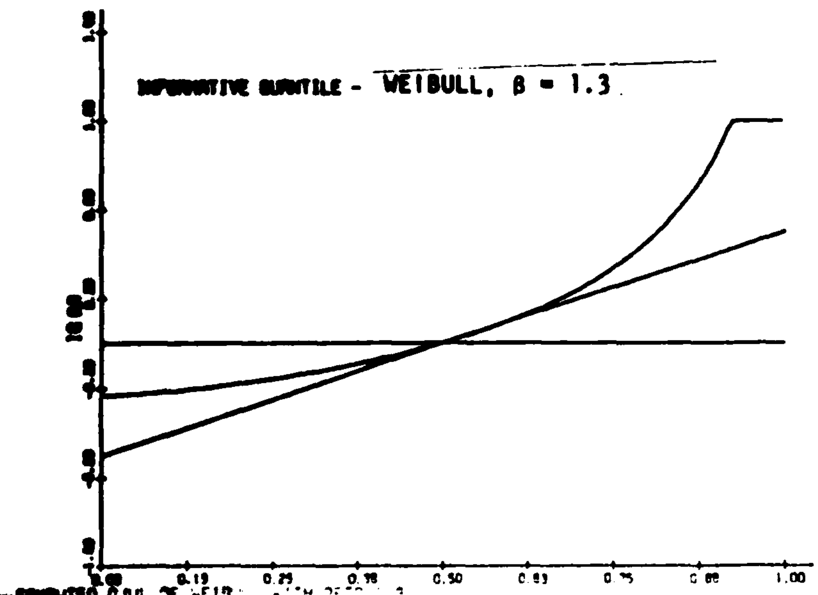

INFORMATIVE QUANTILE - WEIBULL, β = 1.2
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=1.2

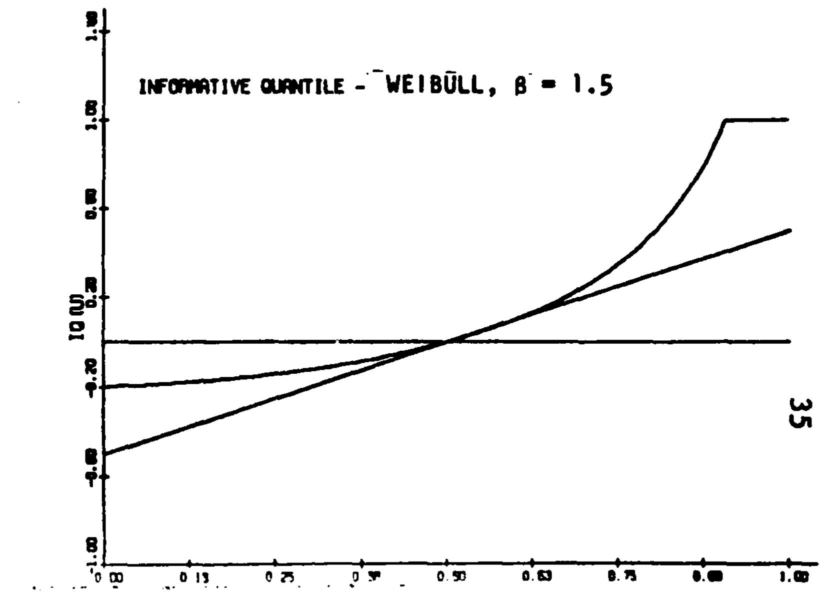

INFORMATIVE QUANTILE - WEIBULL, β = .9
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=.9

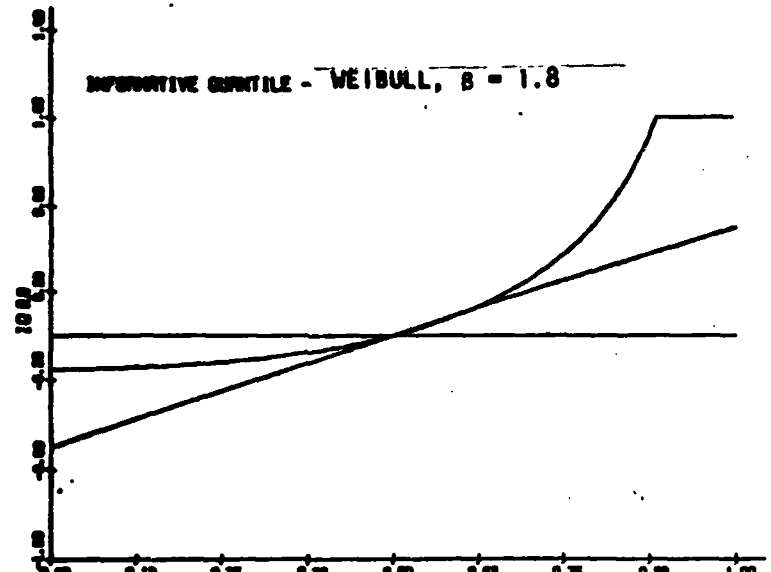

INFORMATIVE QUANTILE - WEIBULL, β = 1.1
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=1.1


INFORMATIVE QUANTILE - WEIBULL, β = 1.4
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=1.4

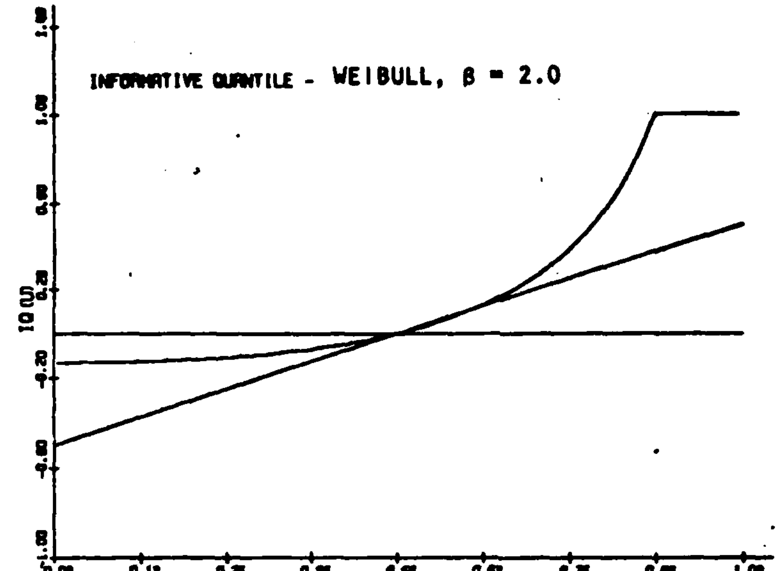

INFORMATIVE QUANTILE - WEIBULL, β = 1.6
IQ(U)
COMPUTED Q(U) OF WEIBULL WITH BETA=1.6


INFORMATIVE QUANTILE - WEIBULL, β = 1.3
IQ(U)

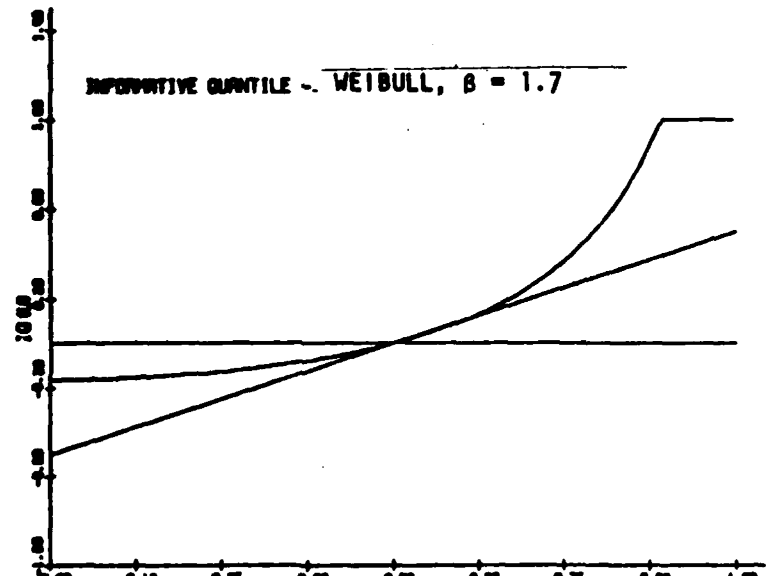

INFORMATIVE QUANTILE - WEIBULL, β = 1.5
IQ(U)



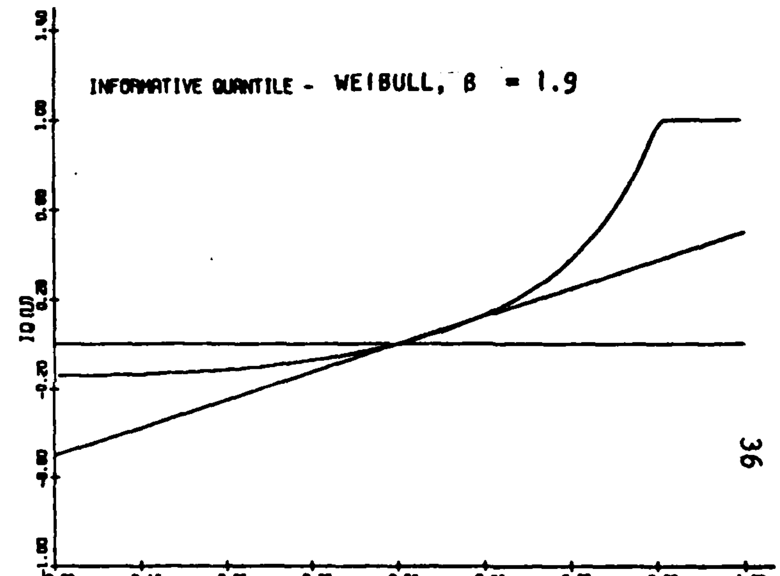
COMPUTED Q(U) OF WEIBULL WITH BETA=1.8



COMPUTED Q(U) OF WEIBULL WITH BETA=2.0



COMPUTED Q(U) OF WEIBULL WITH BETA=1.7



COMPUTED Q(U) OF WEIBULL WITH BETA=1.9

DA

FIL